CH5350: Applied Time-Series Analysis

Method of Moments and Least Squares Estimators

Arun K. Tangirala

Department of Chemical Engineering, IIT Madras

Four broad different methods

The four broad different methods differ mainly in their approach and philosophy.

Method	Data assumption	Estimation approach
MoM	Data is generated by probabilistic process. No assumptions on distribution.	**Sample moments** satisfy the same equations as theoretical ones
LS	For estimation, no assumptions exist on data; To guarantee "good" estimates, a few assumptions are made	**Minimize squared distance** between predictions and observations
MLE	Data is generated by probabilistic process and a known distribution	**Maximize likelihood** of obtaining given data.
Bayesian	Same as in MLE & parameters are uncertain with known **prior** distribution	**Compute posterior distribution** of parameters and use mean, median or mode as estimate.

Method of Moments

Based on the moments of the joint density function of the observations.

Basic Idea

Develop relationships between the parameters and theoretical moments. Assume that sample versions of the moments also satisfy these relations.

MoM . . . contd.

Given N observations $\{y[1], \cdots, y[N]\}$, where $y[k]$ is described by a distribution $F(y[k]; \boldsymbol{\theta})$, we set up the equations

$$g_i(\boldsymbol{\theta}) = M_i(f) = E(Y^i) = \int y^i \, dF = \int y^i f(y) \, dy \qquad i = 1, \cdots, p \qquad (1)$$

where $g_i(\theta)$ is an i^{th} function of the parameters and $M_i(f)$ is the i^{th} moment.

Set up as many equations as the parameters and substitute theoretical moments with sample moments.

Example: Estimation of mean by MoM

Mean

This is the simplest example. The parameter of interest is $\theta = \mu$. We choose the first moment to set up the relation.

Theoretical relation: $\mu = M_1(f) = E(Y) = \int y f(y)\, dy$

Sample version:

$$\hat{\mu} = \frac{1}{N} \sum_{k=0}^{N-1} y[k] \tag{2}$$

which is obtained by replacing the ensemble average with the sample average.

The estimator is the sample mean!

Example: Estimation of mean and variance by MoM

Now apply the MoM to estimate mean and variance of a random process from N observations $\mathbf{y}_N$.

Mean and variance

Here we have two parameters of interest $\theta = \begin{bmatrix} \mu & \sigma^2 \end{bmatrix}^T$. To set up two equations, we use the first two moments of the p.d.f.

Theoretical relations:

$$\mu = M_1(f) = \int y f(y)\, dy; \qquad \sigma^2 + \mu^2 = M_2(f) = \int y^2 f(y)\, dy \tag{3}$$

Example . . . contd.

Sample version: Replace the theoretical averages with sample averages

$$\hat{\mu} = \frac{1}{N}\sum_{k=0}^{N-1} y[k] \qquad \hat{\sigma}^2 + \hat{\mu}^2 = \frac{1}{N}\sum_{k=0}^{N-1} y^2[k] \tag{4}$$

Solving for variance, we obtain

$$\boxed{\hat{\sigma}_y^2 = \frac{1}{N}\sum_{k=0}^{N-1}(y[k] - \bar{y})^2} \qquad \text{(Sample variance)} \tag{5}$$

Example 3: Estimating model parameter

Linear regression

Consider fitting a straight line with no intercept

$$y[k] = \alpha x[k] + \varepsilon[k] \tag{6}$$

where $\varepsilon[k]$ is zero-mean and known variance σ_ε^2. Estimate α given N observations of $y[k]$ and $x[k]$.

Solution: A single parameter is to be estimated. Therefore, we set up a single equation. We obtain two different estimates depending on whether we use first or second moment

Example 3: Solution . . . contd.

1. Writing the first moment equation, we have

$$E(y[k]) = \alpha E(x[k]) + E(\varepsilon[k]) \Longrightarrow \mu_y = \alpha \mu_x \tag{7}$$

Therefore, an estimate is simply

$$\boxed{\hat{\alpha}^\star = \frac{\hat{\mu}_y}{\hat{\mu}_x} = \frac{\bar{y}}{\bar{x}}} \tag{8}$$

Of course, it is required that $x[k]$ be of non-zero mean

Example 3: Solution . . . contd.

2. Writing the second moment equation, we have

$$E((y[k] - \mu_y)^2) = \alpha E((x[k] - \mu_x)^2) + E((\varepsilon[k] - \mu_x \varepsilon)^2) \Longrightarrow \sigma_y^2 = \alpha^2 \sigma_x^2 + \sigma_\varepsilon^2 \tag{9}$$

Therefore, an MoM estimate is

$$\boxed{\hat{\alpha}^\star = \sqrt{\frac{\hat{\sigma}_y^2 - \hat{\sigma}_\varepsilon^2}{\hat{\sigma}_x^2}}} \tag{10}$$

There exists a third way of solving for α!

MoM: Remarks

- The method of moments does not give rise to a unique estimator. A change of moment conditions produces a different estimator.
- MoM estimators built from certain moment conditions, as we shall observe later, are identical to least squares estimators.
- In the identification literature, MoM leads to the so-called *correlation methods*.
- A common application of MoM is in the estimation of time-series models, specifically the AR models, which results in the popular Yule-Walker method.
- MoM estimators are generally used to initialize other estimation methods because by themselves they are known to produce inefficient estimates.

Generalized MoM

In the early 1980s and the years to follow, several researchers (Hall, 2005; Hansen, 1982) generalized the ideas in MoM to arrive at generalized method of moments (GMM) estimators. These are increasingly used in econometrics (Ogaki, 1993). GMM can be used to handle a variety of situations, such as uncertainties in both explanatory variables and measurements, partial knowledge of the model, more moment conditions than parameters, etc.

Further, it can be shown that under certain conditions, GMM is equivalent to the LS and MLE methods. These estimators are known to be asymptotically consistent.

A method belonging to this class that is popularly used in identification is the **instrumental variable method** (and its extended version).

Least Squares Methods

The idea underlying the method of least squares , as intuitive as it appears today, was first published at least two centuries ago. The grass roots of this method are in the fields of astronomy, where a data-driven approach using the least-squares principle was used to predict the trajectory of an asteroid.

Numerous variants and generalizations of the least-squares approach have been proposed, studied and applied. Today it is an indispensable tool in all data-driven approaches to prediction, modelling, control, monitoring, *etc*.

Least Squares Methods

The LS method can be presented in different ways depending on the context. *We shall present it as a method to obtain the best functional approximation for a given signal or a vector*.

We study the **generic problem**, known as regression in statistics. Adaptations to specific estimation problems shall be discussed at appropriate places.

Sample LS problem

The LS method can be formulated for the case where observations of signals are available (known as the **sample LS problem**) or in terms of the actual signals itself (known as the **theoretical LS problem**). We shall study the former.

Problem Statement

Given N observations of a variable $\mathbf{y} = \begin{bmatrix} y[0] & \cdots & y[N-1] \end{bmatrix}$, obtain the best prediction (or approximation) of $\mathbf{y}$ using m *explanatory variables* (or *regressors*) $\varphi_i[k],\ i = 1, \cdots, p$ such that the predictions $\hat{\mathbf{y}}[k]$ are *collectively at a minimum distance* from $\mathbf{y}$.

Remarks

- All predictions should be collectively at a minimum distance, not merely a single prediction from the respective observations.
- The *explanatory variables* φ's could be any known variable that is believed to contain *information* on y

Linear (Ordinary) least squares

The statement of the problem is complete only with a specification of "how" the predictions are made. The entity that does this job is called a **model**. When a linear model is used, we have the **linear least squares** problem.

The prediction (approximation) equation from a linear model is given by,

$$\hat{y}[k] = \sum_{i=1}^{p} \theta_i \varphi_i[k] = \boldsymbol{\varphi}^T[k]\boldsymbol{\theta} \tag{11}$$

where $\boldsymbol{\theta}$ is the vector of parameters that have to be optimized.

Linear (Ordinary) least squares

Introduce the quantities

$$\Phi = \begin{bmatrix} \boldsymbol{\varphi}[0] & \boldsymbol{\varphi}[1] & \cdots & \boldsymbol{\varphi}[N-1] \end{bmatrix}^T; \qquad \mathbf{Z} = \mathbf{y} \cup \Phi \tag{12}$$

The optimization problem can thus be written

$$\min_{\boldsymbol{\theta}} J_N(\mathbf{Z}, \boldsymbol{\theta}) = ||\mathbf{y} - \hat{\mathbf{y}}||_2^2 = (\mathbf{y} - \hat{\mathbf{y}})^T(\mathbf{y} - \hat{\mathbf{y}})$$
$$\text{s.t. } \hat{\mathbf{y}} = \Phi\boldsymbol{\theta}$$

The OLS problem can be solved in different ways. We shall make use of the well-known **projection theorem** in approximation theory to solve the problem.

Solution to the OLS problem

Theorem (Projection Theorem)

Let C be a closed subspace of the Hilbert space $\mathcal{H}$ and let $\mathbf{y}$ be an element in $\mathcal{H}$. Then, $\mathbf{y}$ can be uniquely decomposed into two parts

$$\mathbf{y} = \hat{\mathbf{y}} + \boldsymbol{\varepsilon} \tag{13}$$

where $\hat{\mathbf{y}}$ belongs to C and $\boldsymbol{\varepsilon}$ is orthogonal to C, i.e., $\boldsymbol{\varepsilon}$ is orthogonal to every regressor and, therefore, $\hat{\mathbf{y}}$. This decomposition is unique in the sense that it minimizes the distance between $\mathbf{y}$ and any other vector $\mathbf{w}$ in C,

$$||\mathbf{y} - \hat{\mathbf{y}}||_2 \leq ||\mathbf{y} - \mathbf{w}||_2 \qquad \forall\, \mathbf{w} \in C \tag{14}$$

with the equality holding only when $\mathbf{w} = \hat{\mathbf{y}}$.

Remarks

- The vector $\hat{\mathbf{y}}$ is said to be the **projection** of $\mathbf{y}$ onto the C (or the bases spanning C) and is denoted by $P_C\mathbf{y}$.
- The essence of projection theorem is that **the residuals of a Euclidean distance minimization approach are orthogonal to the explanatory variables (inner products are zero)**.

Solving the OLS problem

In order to apply the projection theorem, recognize that the basis vectors are the regressors and that the residuals are given by $\boldsymbol{\varepsilon} = \mathbf{y} - \hat{\mathbf{y}}$. Then, by virtue of the theorem,

$$\langle \boldsymbol{\varphi}_i, \boldsymbol{\varepsilon} \rangle = 0 \Longrightarrow \boldsymbol{\varphi}_i^T(\mathbf{y} - \Phi\boldsymbol{\theta}) = 0 \qquad i = 1, \cdots, p \tag{15}$$

All the p equations can be jointly written as

$$\Phi^T(\mathbf{y} - \Phi\boldsymbol{\theta}) = \mathbf{0}$$

yielding the familiar solution

$$\boxed{\hat{\boldsymbol{\theta}}_{\mathsf{LS}}^{\star} = (\Phi^T\Phi)^{-1}\Phi^T\mathbf{y}} \tag{16}$$

Predictions and Residuals

The optimal prediction of $\mathbf{y}$ and the associated residuals are:

$$\hat{\mathbf{y}}_{\mathsf{LS}} = \Phi\hat{\boldsymbol{\theta}}_{\mathsf{LS}}^{\star} = \Phi(\Phi^T\Phi)^{-1}\Phi^T\mathbf{y} = \mathbf{P}\mathbf{y} \tag{17}$$

$$\boldsymbol{\varepsilon}_{\mathsf{LS}} = \mathbf{y} - \hat{\mathbf{y}}_{\mathsf{LS}} = (\mathbf{I} - \Phi(\Phi^T\Phi)^{-1}\Phi^T)\mathbf{y} = \mathbf{P}^{\perp}\mathbf{y} \tag{18}$$

where

$$\mathbf{P} = \Phi(\Phi^T\Phi)^{-1}\Phi^T \quad \text{and} \quad \mathbf{P}^{\perp} = \mathbf{I} - \mathbf{P} \tag{19}$$

are said to be the *projection matrix* and its *orthogonal complement* respectively. The latter name is due to the fact that

$$\mathbf{P}\mathbf{P}^{\perp} = \mathbf{P}^{\perp}\mathbf{P} = \mathbf{0} \tag{20}$$

Matrix inversion perspective

The LS estimator can be thought of as a solution to a set of linear *overdetermined* equations since p unknowns are being estimated from N equations while forcing the prediction $\hat{\mathbf{y}}$ to match $\mathbf{y}$

$$\mathbf{y} \approx \Phi\boldsymbol{\theta} \tag{21}$$

Pre-multiplying both sides by Φ^T yields

$$\Phi^T\mathbf{y} = (\Phi^T\Phi)\boldsymbol{\theta} \tag{22}$$

These are known as the *normal equations*.

Pseudo-inverse

Now introduce the *pseudo-inverse*

$$\Phi^\dagger = (\Phi^T\Phi)^{-1}\Phi^T \tag{23}$$

so that the LS solution can be written as

$$\boxed{\hat{\boldsymbol{\theta}}^\star_{\text{LS}} = \Phi^\dagger\mathbf{y}} \tag{24}$$

- Strictly speaking, we are solving $\mathbf{y} = \Phi\boldsymbol{\theta} + \boldsymbol{\varepsilon}$. Imposing $\Phi^T\boldsymbol{\varepsilon} = \mathbf{0}$ eliminates the errors and also sets up the exact set of equations in (22).
- The reason for terming $\Phi^\dagger$ as the **pseudo-inverse** is apparent from a comparison with the case of exact set of equations.

Equivalence of OLS with MoM

From the projection theorem, it is somewhat evident that the OLS estimate of $\boldsymbol{\theta}$ in the linear regression is equivalent to the method of moments estimate.

The equivalence comes from the theoretical LS estimator:

$$\boxed{\Sigma_{\boldsymbol{\varphi}\boldsymbol{\varphi}}\boldsymbol{\theta} = \Sigma_{\boldsymbol{\varphi}Y} \quad \Longrightarrow \quad \boldsymbol{\theta}^{\star}_{\mathsf{LS}} = \Sigma^{-1}_{\boldsymbol{\varphi}\boldsymbol{\varphi}}\Sigma_{\boldsymbol{\varphi}Y}} \tag{25}$$

where the Σ stands for covariance matrix.

Replacing the theoretical covariances by the sample versions gives rise to the sample LS estimator in (16) **as well as the MoM estimator.**

Including a constant term

A constant term (intercept term, or a non-zero mean) can be accommodated in the regression model by simply appending the regressor with a vector of ones as seen below.

$$y[k] = \boldsymbol{\varphi}^T[k]\boldsymbol{\theta} + \beta = \begin{bmatrix}\boldsymbol{\varphi}[k] & 1\end{bmatrix}^T \begin{bmatrix}\boldsymbol{\theta} \\ \beta\end{bmatrix} \tag{26}$$

Interestingly, the LS estimate of the constant term β can be obtained sequentially by first obtaining $\hat{\boldsymbol{\theta}}_{\mathsf{LS}}$ followed by,

$$\boxed{\hat{\beta}_{\mathsf{LS}} = \bar{y} - \bar{\boldsymbol{\varphi}}^T\hat{\theta}_{\mathsf{LS}}} \tag{27}$$

where $\bar{y}$ and $\bar{\boldsymbol{\varphi}}$ are the sample means of $y[k]$ and the regressors respectively.

Goodness of LS fits

Having estimated the parameters, we would like to assess how good the estimated model is? Specifically,

1. **How well has the model (of the specified structure) explained (predicted) the output?** For a given structure and data, the quality of prediction solely depends on the estimation algorithm. We shall focus on this question at present, beginning with the popular R^2 measure.
2. **Is there a need to refine the chosen model structure?** This is a broader question, which requires the use of model diagnostic measures, specifically pertaining to *residual analysis*. The primary tools are cross-correlation of residuals with inputs, auto-correlation of residuals, and cross-validation.

R^2 measure

The R^2 measure is a goodness-of-fit index. It gives a bird's eye view of how well the model has explained the variations in the data. Its definition is based on an important feature of LS estimation:

$$\underbrace{\sum_{k=0}^{N-1}(y[k]-\bar{y})^2}_{\text{sum square total (SST)}} = \underbrace{\sum_{k=0}^{N-1}(\hat{y}[k]-\bar{y})^2}_{\text{sum square predictions (SSP)}} + \underbrace{\sum_{k=0}^{N-1}\varepsilon^2[k]}_{\text{sum square errors (SSE)}} \tag{28}$$

R^2 measure . . . contd.

The total variance of the output is broken up into two additive terms - the variance explained by the model and the variance of the residuals.

Coefficient of determination R^2

$$R^2 \triangleq \frac{SSP}{SST} = 1 - \frac{\sum_{k=0}^{N} \varepsilon^2[k]}{\sum_{k=0}^{N-1} (y[k] - \bar{y})^2} = 1 - \frac{||\hat{\mathbf{y}} - \mathbf{y}||_2^2}{||\mathbf{y} - \bar{y}||_2^2} \tag{29}$$

Remarks on the use of R^2 measure

1. From (29) and (28) it follows that $0 \leq R^2 \leq 1$.
2. The concept of R^2 can also be used with other model fitting methods. However, if the method does not fulfill (28), R^2 can be greater than unity.
3. Interestingly, it can be shown that

$$R^2 = \text{corr}^2(y[k], \hat{y}[k]) \tag{30}$$

4. *It is necessary to include a constant term in the model (one in the regression vectors) to correctly compute R^2 as a measure of fit.*
5. The R^2 metric does not reveal anything about the unexplained portion *i.e.*, whether the residuals carry any predictable characteristics.

Adjusted R^2

6. R^2 has a poor sensitivity with respect to inclusion (or exclusion) of additional regressors. Thus, it cannot be used to determine overfits.
7. An **adjusted** R^2 that is based on the *mean square* is introduced for this purpose:

$$\bar{R}^2 = 1 - \frac{SSE/(N-p)}{SST/N-1} = 1 - \frac{N-1}{N-p}(1-R^2) \tag{31}$$

   The factors $(N-1)$ and $(N-p)$ denote the d.o.f. of SST and SSE, respectively.
8. The modified measure can assume negative values unlike the classical R^2.
9. It measures the balance between prediction bias and the variability of estimates.
10. In practice, sophisticated measures based on information theory such as **AIC** and **SIC** are employed.

Properties of LS estimator

In order to evaluate the properties of any parameter estimator, we first assume a description for the "true" process that generates the measurements, known as the **data generating process** (DGP).

DGP for linear regression

The process is assumed to be linear with additive noise

$$\text{DGP:} \qquad y[k] = \boldsymbol{\varphi}^T[k]\boldsymbol{\theta}_0 + \xi[k] \tag{32}$$

where $\boldsymbol{\theta}_0$ is the **true** parameter vector, $\boldsymbol{\varphi}^T[k]$ is the regressor and $\xi[k]$ contains the *unobserved* stochastic terms that collectively represents the effects of unmeasured disturbances and noise. It is also conventional to call $\xi[k]$ as the *equation error*.

Prediction error

The *prediction error* or the *residual* incurred in using a LS estimate of $\boldsymbol{\theta}$ is

$$\boxed{\varepsilon[k] = y[k] - \hat{y}[k] = y[k] - \boldsymbol{\varphi}^T[k]\hat{\boldsymbol{\theta}} = \boldsymbol{\varphi}^T[k]\tilde{\boldsymbol{\theta}} + \xi[k]} \tag{33}$$

where $\tilde{\boldsymbol{\theta}}$ is the parameter estimation error, given by

$$\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0 - (\Phi^T\Phi)^{-1}\Phi^T(\Phi\boldsymbol{\theta}_0 + \boldsymbol{\xi}) = (\Phi^T\Phi)^{-1}\Phi^T\boldsymbol{\xi} \tag{34}$$

- Observe that **the prediction error for a given data is never equal to the equation error** $\xi[k]$ but additionally contains contributions from the "difference" between the true and estimated parameters.

Properties of the OLS estimator

The properties are listed without proofs. For most of the properties listed, two different cases for Φ, namely, *deterministic* and *stochastic* are considered.

1. **Bias:**
   - **Deterministic** Φ**:** The estimator is unbiased if $E(\xi[k]) = 0$.
   - **Stochastic** Φ**:** The LS estimator is *unbiased whenever the noise term* $\xi[k]$ *in the process is uncorrelated to the regressors*.
     To understand the above result, recall that $E(\tilde{\boldsymbol{\theta}}) = E_\Phi(E(\tilde{\boldsymbol{\theta}}|\Phi))$.

$$E(\tilde{\boldsymbol{\theta}}|\Phi) = E((\Phi^T\Phi)^{-1}\Phi^T\boldsymbol{\xi}|\Phi) = (\Phi^T\Phi)^{-1}\Phi^T E(\boldsymbol{\xi}|\Phi) \tag{35}$$

$$\text{Therefore,} \quad E(\boldsymbol{\xi}|\Phi) = 0 \Longrightarrow E(\tilde{\boldsymbol{\theta}}|\Phi) = 0 \Longrightarrow E(\tilde{\boldsymbol{\theta}}) = 0 \tag{36}$$

Properties of the OLS estimator

2. **Variance:** Remember that we are estimating p parameters and hence we speak of a variance-covariance matrix $\Sigma_{\hat{\boldsymbol{\theta}}}$. Assuming unbiasedness, variance $\equiv$ MSE.

$$\Sigma_{\hat{\boldsymbol{\theta}}|\Phi} = E((\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T|\Phi) = (\Phi^T\Phi)^{-1}\Phi^T\Sigma_{\boldsymbol{\xi}}\Phi(\Phi^T\Phi)^{-1} \tag{37}$$

   - **White observation errors:** $\Sigma_{\boldsymbol{\xi}} = \sigma_e^2\mathbf{I}_{N\times N}$.

$$\boxed{\Sigma_{\hat{\boldsymbol{\theta}}} = \sigma_e^2(\Phi^T\Phi)^{-1}} \tag{38}$$

   - The variance above is the Cramer-Rao bound for all unbiased estimators of the linear regression model - therefore OLS is **efficient when $\xi[k]$ is white**.

Estimation of variance

- To be able to use (38), we need an estimate of σ_e^2. An unbiased estimator is:

$$\boxed{\hat{\sigma}_e^2 = \frac{\sum_{k=0}^{N-1}\varepsilon^2[k]}{N-p} = \frac{\text{SSE}}{N-p}} \qquad \text{Note: The SSE has } (N-p) \text{ d.o.f.} \tag{39}$$

- Further,

$$E(\hat{\sigma}_e^2) = \sigma_e^2 \qquad \text{(Unbiased)} \tag{40a}$$

$$E((\hat{\sigma}_e^2 - \sigma_e^2)^2) = \frac{2\sigma_e^2}{N-p} \qquad \text{(Consistent)} \tag{40b}$$

$$(N-p)\frac{\hat{\sigma}_e^2}{\sigma_e^2} \xrightarrow{d} \chi^2_{N-p} \tag{40c}$$

Efficiency of OLS estimator . . . contd.

3. **Efficiency:** As remarked earlier, the OLS estimator has the lowest variance among all estimators of the linear regression model **when the equation error $\xi[k]$ of the generating process is GWN**.

   A variant of the LS technique known as the *weighted* least squares (WLS) (to be discussed shortly), produces efficient estimates even when $\xi[k]$ is coloured.

Consistency of the OLS estimator

4. **Consistency:** The OLS estimates converge (in the sense of probability) to the true value provided
   i. The covariance of regressors is $E(\boldsymbol{\varphi}[k]\boldsymbol{\varphi}^T[k]) = \Sigma_{\boldsymbol{\varphi}\boldsymbol{\varphi}}$ is invertible
   ii. The regressors are uncorrelated with equation errors, $E(\boldsymbol{\varphi}[k]\xi[k]) = \mathbf{0}$.

   Mean square consistency is guaranteed when $\xi[k]$ is white with deterministic Φ.

Distribution of the OLS estimates

5. **Distribution:** The requirements depend on the nature of errors, i.e., whether they are Gaussian or not, but the end result remains the same.
   i. *Gaussian errors*: The conditional distribution of the estimates is Gaussian. This is easy to establish, because a linear combination of Gaussian distributed RVs produces a RV with Gaussian distribution.

$$\hat{\boldsymbol{\theta}} = \Phi^{\dagger}\mathbf{y} \qquad \boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \Sigma_{\boldsymbol{\xi}}) \Longrightarrow \hat{\boldsymbol{\theta}}|\Phi \sim \mathcal{N}(\boldsymbol{\theta}_0, \sigma_e^2(\Phi^T\Phi)^{-1}) \tag{41}$$

   ii. *Non-Gaussian errors:* In this case too, the estimates follow a Gaussian distribution by virtue of the CLT. This holds even when the equation error is coloured and for deterministic / stochastic regressors. However, this is an *asymptotic* property.

Distribution of OLS estimates

If $\xi[k]$ are **independent** and identically distributed (i.i.d.) with mean zero and variance σ^2 and the regressors are "well-behaved", then

$$\hat{\boldsymbol{\theta}} \xrightarrow{d} \mathcal{N}\left(\boldsymbol{\theta}_0, \frac{\sigma^2}{N}\Sigma_{\varphi\varphi}^{-1}\right) \tag{42}$$

- By well-behaved regressors it is meant that
  (i) $\Phi^T\Phi$ is of full rank as $N \to \infty$
  (ii) No single observation shall dominate the data.
- In practice, the distribution properties are computed by replacing the theoretical quantities with their corresponding sample versions,

Distribution of a function of parameters

In many situations, the actual parameters of interest may be a function $\mathbf{g}(\boldsymbol{\theta})$.

Distribution of a function of estimates

If $\mathbf{g}(\hat{\boldsymbol{\theta}})$ is a set of continuous and continuously differentiable functions of $\hat{\boldsymbol{\theta}}$, then under the conditions of Theorem 40,

$$\mathbf{g}(\hat{\boldsymbol{\theta}}) \xrightarrow{d} \mathcal{N}\left(\mathbf{g}(\boldsymbol{\theta}_0), \boldsymbol{\Omega}\left(\frac{\sigma^2}{N}\Sigma_{\varphi\varphi}^{-1}\right)\boldsymbol{\Omega}^T\right) \tag{44}$$

where $\boldsymbol{\Omega} = d\mathbf{g}/d\boldsymbol{\theta}$ is the Jacobian of the functions $\mathbf{g}$ w.r.t. the parameter vector.

- In practice, the theoretical variance is replaced by the asy. version: $\hat{\boldsymbol{\Omega}}(\hat{\sigma}^2(\Phi^T\Phi)^{-1})\hat{\boldsymbol{\Omega}}^T$

Confidence Intervals from LS estimates

When the conditions for (42) are met, the (standardized) individual parameter estimates have a standard normal distribution,

$$\boxed{\frac{\hat{\theta}_i - \theta_{i0}}{\sqrt{\sigma_e^2 S_{ii}}} \sim \mathcal{N}(0,1)} \tag{45}$$

where S_{ii} is the i^{th} diagonal element of $(\Phi^T\Phi)^{-1}$.

- When σ_e^2 is replaced by its estimator in (39), $\hat{\theta}_i$ has a Student's t-distribution,
- The $100(1-\alpha)\%$ confidence interval for θ_{i0} is therefore,

$$\boxed{\hat{\theta}_i - t_{1-\alpha/2,N-p}\sqrt{\hat{\sigma}_e^2 S_{ii}} \leq \theta_{i0} \leq \hat{\theta}_i + t_{1-\alpha/2,N-p}\sqrt{\hat{\sigma}_e^2 S_{ii}}} \tag{46}$$

Remarks

- In deriving the properties of the LS estimator, we have assumed that **the functional form of the process has been "rightly" specified.** In practice, this rarely holds since the real process is far more complex than the one in (32).
  - In practice, we turn these requirements to as that of obtaining **white** residuals, which is tested by applying whiteness tests on residuals.
- The sample size N in all the above expressions should be treated as the size of *effective observations* used for estimation, which depends on the problem in hand.
  - When fitting AR(P) models, one sacrifices the first P observations, thus the effective sample size is $N - P$.

Computing the OLS estimate

The theoretical expression for computing the OLS estimate is not amenable for computation since an inverse is involved.

- In fact,

$$\text{cond}(\Phi^T\Phi) = \text{cond}^2(\Phi)$$

  where cond(.) is the condition number, which makes the normal equations more ill-conditioned than the rectangular set of equations.

- Problems are compounded when $\Phi^T\Phi$ is singular, at least up to working numerical precision.

To circumvent these problems, two numerically efficient methods are widely used. In both methods, the idea is to estimate in a "transformed" space.

QR factorization for OLS

1. **QR factorization:** A QR factorization of Φ is carried out.

$$\Phi \mathbf{P} = \mathbf{Q} \begin{pmatrix} \mathbf{R}_1 \\ \mathbf{0} \end{pmatrix} \tag{48}$$

where the orthogonal *permutation* matrix $\mathbf{P}$ is $p \times p$, orthogonal factor $\mathbf{Q}$ is $N \times N$ and the upper triangular matrix $\mathbf{R}_1$ is $p \times p$. Since Φ is full rank, $\mathbf{R}_1$ is non-singular.

Observations are projected on to the $\mathbf{Q}$ space and partitioned. $\mathbf{Q}^T\mathbf{y} = \begin{pmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \end{pmatrix}$ where $\mathbf{z}_1 : \mathbb{R}^{p \times p}$ and $\mathbf{z}_2 : \mathbb{R}^{N-p \times p}$.

The solution can be shown to be,

$$\boxed{\boldsymbol{\theta} = (\Phi^T\Phi)^{-1}\Phi^T\mathbf{y} = \mathbf{R}_1^{-1}\mathbf{z}_1;} \quad \text{where} \quad \bar{\boldsymbol{\theta}} = \mathbf{P}\boldsymbol{\theta} \tag{49}$$

SVD for OLS

2. **SVD:** In this method, a SVD factorization of the regressor matrix, $\Phi = \mathbf{U}\mathbf{S}\mathbf{V}^T$ is performed. Observations and parameters are projected on to the $\mathbf{U}$ and $\mathbf{V}$ spaces respectively.

$$(\mathbf{y} - \Phi\boldsymbol{\theta})^T(\mathbf{y} - \Phi\boldsymbol{\theta}) = \sum_{i=1}^{r}(\sigma_i\xi_i - \mathbf{u}_i^T\mathbf{y})^2 + \sum_{i=r+1}^{N}(\mathbf{u}_i^T\mathbf{y})^2; \qquad \xi \triangleq \mathbf{V}^T\boldsymbol{\theta} \tag{50}$$

If Φ is of rank r (rank deficient), $(p-r)$ *transformed* parameters are set arbitrarily

$$\xi_i = \begin{cases} \dfrac{\mathbf{u}_i^T\mathbf{y}}{\sigma_i}, & i = 1, \cdots, r \\ \text{arbitrary}, & i = r+1, \cdots, p \end{cases} \tag{51}$$

SVD for OLS contd.

A (1-norm) minimizing solution is to set the "free" $\xi_i = 0,\ i = r+1, \cdots, p$. The resulting optimal parameter estimates $\hat{\boldsymbol{\theta}}$ are,

$$\boldsymbol{\theta} = \mathbf{V}\boldsymbol{\xi} = \sum_{i=1}^{r} \frac{\mathbf{u}_i^T \mathbf{y}}{\sigma_i} \mathbf{v}_i \tag{52}$$

This is also the solution given by **pseudo-inverse**.

MATLAB: `pinv, svd`

Summary

- **Method of Moments** provides elementary means of estimating parameters by developing a map of $\boldsymbol{\theta}$ to the theoretical moments $\mathcal{M}(f(\mathbf{y}))$.
  - $f(\mathbf{y})$ not required. Theoretical moments are replaced by sample versions.
  - Not unique and usually give inefficient estimates
  - Generalized MoM can give consistent and efficient estimates.

Summary . . . contd/

- **Least Squares** estimators optimize parameters such that they minimize the squared distance between $\hat{\mathbf{y}}$ and $\mathbf{y}$.
  - Ubiquitous in estimation. Conceptually simple and have several nice interpretations
  - With linear predictors, unique solutions are obtained
  - Solution does not require any probabilistic assumptions on data
  - Produces asymptotically consistent and normally distributed estimates.
  - Give efficient estimates of linear models when observation errors are white
  - Computations carried out using QR and SVD factorizations.

Variants

There exist several variants of least squares methods:

- **Weighted (Generalized) least squares**
- Partial least squares
- Total least squares
- **Non-linear least squares**
- . . .

Weighted LS: Motivation

A useful generalization of OLS ascribes different level of "importance" to each observation by means of weights $w[k]$. There are several compelling reasons for such a formulation:

1. Error in each observation has a different variance, i.e., *heteroskedastic* errors. Samples with "more" errors in them are less reliable than those with "less" errors (e..g., error changes with operating conditions, data containing outliers, etc.)
2. Observation errors are coloured (recall that OLS is efficient only when $\xi[k]$ is GWN)
3. Recent samples to be given more importance than the past for model adaptation. The weights are then termed as *forgetting factors*.
4. Data obtained from different sensors may have different error characteristics or a set of observations from a single sensor may exhibit different SNR.

WLS

The general statement of the problem is as follows.

$$\min_{\boldsymbol{\theta}} (\mathbf{y} - \Phi\boldsymbol{\theta})^T \mathbf{W} (\mathbf{y} - \Phi\boldsymbol{\theta}) \tag{53}$$

where $\mathbf{W}$ is a *positive definite* **weighting** matrix. By definition therefore, $\mathbf{W}$ is symmetric. When $\mathbf{W} = \mathbf{I}_{N\times N}$, we recover the OLS formulation.

- The cost function in (53) is also known as the *Mahalanobis* distance (Mahalanobis, 1936), and is conventionally written as a weighted squared-norm $||\mathbf{y} - \Phi\boldsymbol{\theta}||_{\mathbf{W}}$.
- The WLS formulation in (53) is also known as the *generalized least squares* (GLS), wherein the case of diagonal $\mathbf{W}$ is the weighted least-squares problem.

The WLS problem can be cast as an OLS problem on scaled data.

Solution to the WLS problem

Since $\mathbf{W}$ is positive-definite, we can perform a Cholesky factorization:

$$\mathbf{W} = \mathbb{C}^T\mathbb{C} \tag{54}$$

Then, the objective function in (53) can be re-written as

$$(\mathbf{y} - \Phi\boldsymbol{\theta})^T\mathbb{C}^T\mathbb{C}(\mathbf{y} - \Phi\boldsymbol{\theta}) \tag{55}$$

Now, introduce scaled observations and regressors,

$$\mathbf{y}_S = \mathbb{C}\mathbf{y};\ \ \Phi_S = \mathbb{C}\Phi \tag{56}$$

The WLS solution

The WLS problem can be then cast into an OLS formulation

$$\min_{\boldsymbol{\theta}} (\mathbf{y}_S - \Phi_S\boldsymbol{\theta})^T(\mathbf{y}_S - \Phi_S\boldsymbol{\theta}) \tag{57}$$

From the OLS solution, we thus have the WLS estimator

$$\boxed{\hat{\boldsymbol{\theta}}_{\text{WLS}} = (\Phi_S^T\Phi_S)^{-1}\Phi_S^T\mathbf{y}_S = (\Phi^T\mathbf{W}\Phi)^{-1}\Phi^T\mathbf{W}\mathbf{y}} \tag{58}$$

- **Scaling the data amounts to scaling the observation errors as well**,

$$\boldsymbol{\xi}_S = \mathbb{C}\boldsymbol{\xi} \tag{59}$$

Remarks

- The scaled errors have the covariance matrix $\Sigma_{\boldsymbol{\xi}_S} = \mathbb{C}\Sigma_{\boldsymbol{\xi}}\mathbb{C}^T$. For a special choice of $\mathbb{C}$ we can render $\Sigma_{\boldsymbol{\xi}_S} = \mathbf{I}$, offering certain advantages.
- Diagonal elements of $\mathbf{W}$ represent the importance given to each observation, while the off-diagonal elements account for the importance given to correlation in the errors. To see this, re-write the objective function as

$$(\mathbf{y} - \Phi^T\boldsymbol{\theta})^T\mathbf{W}(\mathbf{y} - \Phi^T\boldsymbol{\theta}) = \text{trace}((\mathbf{y} - \Phi^T\boldsymbol{\theta})^T\mathbf{W}(\mathbf{y} - \Phi^T\boldsymbol{\theta}))$$

- **Cross-terms in $\mathbf{W}$ handle temporally correlated equation errors.**
- The WLS method is a special case of the GMM and the MLE.

Choice of weights

How to choose the weights matrix $\mathbf{W}$?

The answer depends on the application / criterion.

i. **Model updation:** $\mathbf{W}$ is a diagonal matrix of *forgetting factors*

ii. **Efficient estimates:** The goal here is to achieve minimum $\text{var}(\hat{\boldsymbol{\theta}}_{\text{WLS}})$.

$$\Sigma_{\hat{\boldsymbol{\theta}}}(\mathbf{W}) = \text{var}(\hat{\theta}_{\text{WLS}}) = (\Phi^T\mathbf{W}\Phi)^{-1}\Phi^T\mathbf{W}\Sigma_{\boldsymbol{\xi}}\mathbf{W}\Phi(\Phi^T\mathbf{W}\Phi)^{-1} \tag{60}$$

Choice of weights for efficient estimates

The **optimal weighting matrix is the inverse of the covariance of equation errors**,

$$\boxed{\mathbf{W}_{\text{opt}} = \Sigma_{\boldsymbol{\xi}}^{-1}} \tag{61}$$

With this choice, the variance of the WLS estimator is

$$\boxed{\Sigma_{\hat{\boldsymbol{\theta}}}(\mathbf{W}_{\text{opt}}) = (\Phi^T \mathbf{W}_{\text{opt}} \Phi)^{-1} = (\Phi^T \Sigma_{\boldsymbol{\xi}} \Phi)^{-1}} \tag{62}$$

The result is nicely understood in systems with heteroskedastic errors, as we shall observe next.

Heteroskedastic errors

When $\xi[k]$ is **heteroskedastic** (changing variance),

$$\Sigma_{\boldsymbol{\xi}} = E(\boldsymbol{\xi}\boldsymbol{\xi}^T) = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_N^2 \end{bmatrix} \quad \Longrightarrow \mathbf{W}_{\text{opt}} = \Sigma_{\boldsymbol{\xi}}^{-1} = \begin{bmatrix} 1/\sigma_1^2 & 0 & \cdots & 0 \\ 0 & 1/\sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/\sigma_N^2 \end{bmatrix}$$

The WLS objective function can then be expanded as

$$(\mathbf{y} - \Phi\boldsymbol{\theta})^T \mathbf{W} (\mathbf{y} - \Phi\boldsymbol{\theta}) = \sum_{k=1}^{N} (y[k] - \boldsymbol{\varphi}^T[k]\boldsymbol{\theta})^2 / \sigma_k^2 \tag{63}$$

Higher the error variance, lower is the reliability of that sample, implying **WLS attaches lower importance to those samples with more error.**

Example: Heteroskedastic errors

Sensor fusion: Steady-state estimation

Temperature measurements of a reactor at steady-state from 10 thermocouples that have different, but known error characteristics (variability).

Sensor	1	2	3	4	5	6	7	8	9	10
Meas. $(^\circ C)$	61.2	64.3	59.1	64.1	63.8	62.9	58.2	60.7	61.5	63.7
Variance	0.36	2.25	1.69	0.25	0.49	2.89	3.2	1.4	1.2	2.7

where the readings have already been adjusted for calibration.

Estimate the steady-state temperature from these ten different measurements.

Example . . . contd.

Solution: Assume that the error across sensors are uncorrelated. Then using (61), the optimal weighting is

$$\mathbf{W}_{\text{opt}} = \Sigma_{\boldsymbol{\xi}}^{-1} = \begin{bmatrix} 1/\sigma_1^2 & 0 & \cdots & 0 \\ 0 & 1/\sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/\sigma_{10}^2 \end{bmatrix}$$

Further, note that $\boldsymbol{\varphi}[k] = 1$ for this example.

Example: Heteroskedastic errors ... contd.

The WLS estimate of the average and its error variance are then given by,

$$\hat{\mu}_{WLS} = (\Phi^T \mathbf{W} \Phi)^{-1} \Phi^T \mathbf{W} \mathbf{y} = \frac{\sum_{k=1}^{N} \frac{y[k]}{\sigma_k^2}}{\sum_{k=1}^{N} \frac{1}{\sigma_k^2}} = 62.6086;$$

$$\text{var}(\hat{\mu}) = (\Phi^T \mathbf{W} \Phi)^{-1} = \frac{1}{\sum_{k=1}^{N} \frac{1}{\sigma_k^2}} = 0.0805$$

where k indicates the sensor index for this example.

Example: Heteroskedastic errors ... contd.

Compare corresponding results from OLS

$$\hat{\mu}_{OLS} = \frac{1}{N} \sum_{k=1}^{N} y[k] = 61.95; \quad \text{var}(\hat{\mu}) = (\Phi^T \Phi)^{-1} \Phi^T \Sigma_{\boldsymbol{\xi}} \Phi (\Phi^T \Phi)^{-1} = \frac{\sum_{k=1}^{N} \sigma_k^2}{N^2} = 0.1643 \tag{64}$$

The widths of confidence intervals for the average correspondingly would be proportional to $2\sigma_{\hat{\mu}}$, i.e., 0.2835 and 0.4053 respectively.

Remarks

- WLS is used for estimating parameters of equation error models with coloured observation errors (e.g., ARMA models)
- The optimal weights are generally unknown since $\Sigma_{\boldsymbol{\xi}}$ is not known a priori. Therefore, the general practice is to use an iterative method.
- Alternatively, propose a model for the errors or its variance and jointly estimate the parameters of this model as well as the regression equation. This has strong semblance to likelihood methods.
- The weighted least squares with non-diagonal weights is commonly known as **generalized least squares** method
- Statistical tests for heteroskedasticity are available. Popular among these are White's test, and the Park, Glesjer, and Breusch-Pagan-Godfrey tests.

Non-linear Least Squares

The OLS and its variants discussed until this point were formulated using linear predictors (models). A practically useful generalization is the method of non-linear least squares (NLS), which handles non-linear models.

The NLS problem statement:

$$\min_{\boldsymbol{\theta}} \mathrm{J}_N(\boldsymbol{\theta}, \mathbf{y}, \boldsymbol{\varphi}) = \frac{1}{N}||\mathbf{y} - \hat{\mathbf{y}}(\boldsymbol{\theta}, \boldsymbol{\varphi})||_2^2 \qquad \text{s.t. } \hat{\mathbf{y}}(\boldsymbol{\theta}, \boldsymbol{\varphi}) = \mathbf{s}(\boldsymbol{\theta}, \boldsymbol{\varphi}) \tag{65}$$

where $\mathbf{s}(.)$ is a known (or user-specified) non-linear transformation, $\mathbf{y}$ is the $N \times 1$ observation vector and $\boldsymbol{\varphi}$ is the set of explanatory variables as usual.

Note: For simplicity, we shall use $\hat{\mathbf{y}}$ in place of $\hat{\mathbf{y}}(\boldsymbol{\theta}, \boldsymbol{\varphi})$.

Solution to the NLS

The optimal solution is once again obtained by setting $\nabla_{\boldsymbol{\theta}} \mathsf{J} = 0$:

$$\boldsymbol{\theta}^{\star} = \text{sol}\left[\mathbf{g}(\boldsymbol{\theta}) \triangleq \nabla_{\boldsymbol{\theta}} \mathsf{J} = -\frac{1}{N}\frac{\partial \hat{\mathbf{y}}}{\partial \boldsymbol{\theta}}^{T}(\mathbf{y} - \hat{\mathbf{y}}) = \mathbf{0}\right] \tag{66}$$

- As in OLS, an orthogonality condition governs the optimum
- **No closed-form and unique solution** unlike in OLS
- $\dim(\boldsymbol{\theta}) \neq \dim(\boldsymbol{\varphi})$
- **Only a numerical solution and local optimum can be obtained**

Solution to NLS problem . . . contd.

Several methods are available, all of which make use of an iterative search.

$$\boxed{\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)} - \eta_i \mathbf{d}^{(i)}} \tag{67}$$

where $\mathbf{d}^{(i)}$ is the **direction** of change in the parameter space, and η_i is the **step length** that controls the amount of change.

- Newton-Raphson
- Gauss-Newton
- Steepest descent, Levenberg-Marquardt, Quasi-Newton, Trust region

Newton-Raphson method

The N-R method is based on the following choice of direction and step change:

$$\mathbf{d}^{(i)} = \mathbf{g}_i; \quad \eta_i = (\mathbf{H}^{(i)})^{-1} \tag{68}$$

where $\mathbf{g}_i = (\nabla_{\boldsymbol{\theta}} \mathsf{J})^{(i)}$ is the Jacobian and $\mathbf{H}^{(i)} = \nabla^2 \mathsf{J}_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(i)}}$ is the $p \times p$ Hessian.

Shortcomings of N-R method

- Computation of a matrix inverse and the Hessian is involved at each iteration
- Positive-definiteness of Hessian is not guaranteed, meaning, objective function is not bound to decrease after every iteration.

The *modified N-R method* overcomes these drawbacks by modifying an additional factor in the step length:

$$\boxed{\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)} - \alpha_i (\mathbf{H}^{(i)})^{-1} \mathbf{g}_i} \tag{69}$$

Gauss-Newton method

The Gauss-Newton method employs an OLS on a first-order approximation of the non-linear predictor at each iteration:

$$\hat{\mathbf{y}}(\boldsymbol{\theta}) \approx \hat{\mathbf{y}}(\boldsymbol{\theta}^{(i)}) + \Psi(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(i)}}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(i)}) \tag{70}$$

where Ψ is made up of the gradients of the predictor

$$\boldsymbol{\psi}(k,\boldsymbol{\theta}) \triangleq \nabla_{\boldsymbol{\theta}}\hat{y}(k,\boldsymbol{\theta}) = \frac{\partial \hat{y}(k,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \begin{bmatrix} \frac{\partial \hat{y}[k]}{\partial \theta_1} & \cdots & \frac{\partial \hat{y}[k]}{\partial \theta_p} \end{bmatrix}^T \tag{71}$$

$$\Psi(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\hat{\mathbf{y}} = \begin{bmatrix} \boldsymbol{\psi}(0,\boldsymbol{\theta}) & \boldsymbol{\psi}(1,\boldsymbol{\theta}) & \cdots & \boldsymbol{\psi}(N-1,\boldsymbol{\theta}) \end{bmatrix}^T \tag{72}$$

G-N Method . . . contd.

The choices of step length and search direction for the G-N method are:

$$\mathbf{d}^{(i)} = \Psi(\boldsymbol{\theta}^{(i)})^T(\mathbf{y} - \hat{\mathbf{y}}(\boldsymbol{\theta}^{(i)})) \qquad \eta_i = (\Psi(\boldsymbol{\theta}^{(i)})^T\Psi(\boldsymbol{\theta}^{(i)}))^{-1} \tag{73}$$

- It can be shown that this approach is *equivalent to a N-R method with a suitable approximation of the Hessian*.

$$\begin{aligned} \mathbf{H}(\boldsymbol{\theta}) &= \nabla - \left(\frac{1}{N}\Psi(\boldsymbol{\theta})^T\varepsilon(\boldsymbol{\theta})\right) \\ &= \frac{1}{N}(\Psi(\boldsymbol{\theta})^T\Psi(\boldsymbol{\theta})) - \frac{1}{N}((\nabla_{\boldsymbol{\theta}}\Psi)^T\boldsymbol{\varepsilon}(\boldsymbol{\theta}) \\ &\approx \frac{1}{N}(\Psi(\boldsymbol{\theta})^T\Psi(\boldsymbol{\theta})) \quad \Longrightarrow \eta = \frac{1}{N}\mathbf{H}^{-1} \end{aligned} \tag{74}$$

Gauss-Newton method . . . contd.

- The approximation in (74) can have very slow to zero convergence rate when the residuals are large. To circumvent this problem, a *damped* Gauss-Newton method is often employed:

$$\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)} - \mu_i(\Psi(\boldsymbol{\theta}^{(i)})^T\Psi(\boldsymbol{\theta}^{(i)}))^{-1}\Psi(\boldsymbol{\theta}^{(i)})^T\boldsymbol{\varepsilon}_i \tag{75}$$

  where the damping factor μ_i is adjusted such that J_N decreases with every iteration.

See Bates and Watts, 1988 and Seber and Wild, 1989 for additional reading.

Special cases

Non-linear regressions specialize to simpler problems when the predictors assume certain special forms. Three such situations are discussed below:

1. **Linear in parameters:** In this case, the predictor has the form

$$\hat{y}[k] = f(\boldsymbol{\varphi}[k])\boldsymbol{\theta} \tag{76}$$

   The parameters can be estimated using an OLS algorithm with regressors now as functions of explanatory variables.

Special cases . . . contd.

2. **Linear via transformation:** The non-linear relationship between y and $\boldsymbol{\varphi}$ is linear in a **transformed** domain

A classical example arises in physical chemistry, where the goal is to estimate the Arrhenius constant from kinetics data

$$k_r = e^{-E_a/RT} \quad \Longrightarrow \quad \ln k_r = -\frac{E_a}{RT}$$

- In the transformation approach, it is important to understand that **the error characteristics are also transformed**. In the example above, the errors (in rate constants) are also log-transformed, which can have a serious influence on the properties of the parameter estimates.

Special cases . . . contd.

3. **Pseudo-linear regression:** This is the case when the predictor has a linear regression form, but the regressors are implicit functions of $\boldsymbol{\theta}$

$$\hat{y}[k] = \boldsymbol{\varphi}^T[k, \boldsymbol{\theta}]\boldsymbol{\theta} \tag{77}$$

These forms are widely encountered in the prediction expressions and estimation of ARMA models.

Algorithmic aspects of NLS estimation algorithms

i. **Gradient computation:** Gradients can only be numerically computed in general. Analytical expressions for carrying out these calculations exist only for select predictor forms. Fortunately, the parametric model forms fall into this category.

ii. **Initial guess:** Preliminary guesses of parameters can range from random to estimates from other methods. NLS algorithms can be very sensitive to initial guesses. Therefore, it is always recommended to take any prior information and physical knowledge of the process into account while feeding the initial guess.

iii. **Stopping criteria:** The influence of this factor is relatively less significant on the final estimate. Typical stopping criteria include a combination of tolerance specifications on (a) decrease in J, (b) gradient of objective function $\nabla_{\boldsymbol{\theta}} \mathrm{J}$ and (c) change in $\hat{\boldsymbol{\theta}}$ across two successive iterations.

Asymptotic properties of NLS estimators

Analyzing the properties of NLS estimators is not trivial due to the complexity of the associated algorithms. Presenting these details is beyond the scope of the course. Only the salient results are stated below.

Once again, the data generating process is assumed to be

$$y[k] = s(\boldsymbol{\theta}, \boldsymbol{\varphi}[k]) + \xi[k] \tag{78}$$

Asymptotic properties of NLS . . . contd.

Standard assumptions:

i. *Identifiability*: The requirement is that $s(\boldsymbol{\theta}_1, \boldsymbol{\varphi}) = s(\boldsymbol{\theta}_2, \boldsymbol{\varphi}) \Leftrightarrow \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$.

ii. *Differentiable functional form*: Necessary for the existence of gradients, and even for a solution to exist.

iii. Correlation between gradient and disturbance converges to zero at the optimum.

iv. *Stochastic nature of* $\xi[k]$: The disturbance is *conditionally* zero-mean, homoscedastic, zero temporal correlation and has finite second-order moments.

v. *Explanatory variables are exogenous*: Implies $\text{corr}(\boldsymbol{\varphi}[k], \xi[k]) = 0$.

Consistency

Theorem

Under the conditions of

1. *Compact parameter space: The space* Θ *to which* $\boldsymbol{\theta}$ *belongs is closed and bounded.*
2. *Convergence of the objective function:*

$$\mathrm{J}_N(\boldsymbol{\theta}, \Phi) \xrightarrow{p} \mathrm{J}(\boldsymbol{\theta}) \quad \forall \boldsymbol{\theta} \qquad \textit{(should be continuous and differentiable)}$$

Consistency of NLS estimators ... contd.

3. *Continuity of* $\mathrm{J}(\boldsymbol{\theta})$: The objective function is continuous and differentiable on the parameter space Θ.
4. *Unique minimum of* $\mathrm{J}(\boldsymbol{\theta})$: The obj. fun. $\mathrm{J}(\boldsymbol{\theta})$ has a unique minimum at $\boldsymbol{\theta}_0$.

the LS estimator of the parameters $\boldsymbol{\theta} \in \Theta$ of the non-linear regression model is weakly consistent

$$\hat{\boldsymbol{\theta}}^{\star}_{\mathrm{NLS}} \xrightarrow{p} \boldsymbol{\theta}_0 \tag{79}$$

See Amemiya, 1985 and Greene, 2012 for proofs and further reading.

Asymptotic normality

The NLS estimates asymptotically follow a Gaussian distribution regardless of the actual distribution of the noise term $\xi[k]$, provided the following conditions are met:

i. $\frac{1}{N}\Psi(\boldsymbol{\theta}_0)^T\Psi(\boldsymbol{\theta}_0) \xrightarrow{p} \Sigma^0_{\Psi}$ (positive definite covariance matrix)

ii. $\frac{1}{\sqrt{N}}\Psi(\boldsymbol{\theta}_0)^T\mathbf{v} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \sigma_e^2\Sigma^0_{\Psi})$ (zero correlation between pseudo-regressors and disturbance)

With these assumptions: $\boxed{\hat{\boldsymbol{\theta}}_{\mathrm{NLS}} \sim As\mathcal{N}\left(\boldsymbol{\theta}_0, \frac{\sigma_e^2}{N}(\Sigma^0_{\Psi})^{-1}\right)}$

Estimation of σ_e^2

A consistent estimator of σ_e^2 is given by

$$\boxed{\hat{\sigma}_e^2 = \frac{1}{N}||\mathbf{y} - \hat{\mathbf{y}}(\hat{\boldsymbol{\theta}}, \boldsymbol{\varphi})||_2^2} \tag{80}$$

analogous to the linear LS case.

Note: In practice, the pseudo-regressor is evaluated at the estimated parameter instead of $\boldsymbol{\theta}_0$. See Greene, 2012 for a detailed presentation of this topic.

Remarks

- Observe the strong resemblance between the conditions for optimum, consistency and asymptotic normality of OLS and NLS estimators.
- All results should be expected to carry forward analogously once the similarity of the regressor Φ and pseudo-regressor $\Psi(\boldsymbol{\theta}_0)$ is recognized.
- Hypothesis testing of parameters from the NL regression is more involved. Popular ones include Wald and Lagrange multiplier tests (Greene, 2012; Yan and Su, 2009)
- Small sample behaviour of the NLS estimator is in general different and is difficult to analyze
- Modern ways of computing the statistical properties of the estimate involve the use of **bootstrapping approaches** such as Monte-Carlo simulations and surrogate data methods.

Summary

- Weighted LS estimators extend the use of OLS to the case of auto-correlated and heteroskedastic errors.
- WLS produces efficient estimates when the optimal weighting is the inverse of $\Sigma_{\mathbf{v}}$, the covariance matrix of $\mathbf{v}$
- Non-linear least squares estimators do not have closed-form expressions. They require iterative algorithms, which only yield local minima.
- The gradient of the predictor takes the role of the regressor in the linear LS formulation.
- NLS estimators are asymptotically consistent and efficient under "similar" conditions as those for OLS

Bibliography I

Amemiya, T. (1985). *Advanced Econometrics*. Harvard University Press.

Bates, D. M. and D. G. Watts (1988). *Nonlinear regression: Its applications*. New York, USA: John Wiley & Sons, Inc.

Brockwell, P. and R. Davis (1991). *Time Series: Theory and Methods*. New York, USA: Springer.

Carroll, R. and D. Rupert (1988). *Transformation and Weighting in Regression*. London, UK: Chapman & Hall.

Golub, G. H. and C. F. V. Loan (1996). *Matrix Computations*. 3rd edition. Johns Hopkins University Press.

Greene, W. H. (2012). *Econometric Analysis*. Upper Saddle River, NJ, USA: Prentice Hall.

Hall, A. (2005). *Generalized Method of Moments*. New York, USA: Oxford University Press.

Bibliography II

Hansen, L. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, 50, pp. 1029–1054.

Luenberger, D. (1969). *Optimization by vector space methods*. New York, USA: John Wiley & Sons, Inc.

Mahalanobis, P. C. (1936). On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India*, 2 (1), pp. 49–56.

Ogaki, M. (1993). *Handbook of Statistics*. Vol. 11. Elsevier Science Publishers. Chap. Generalized Method of Moments: Econometric Applications.

Seber, G. and C. Wild (1989). *Nonlinear Regression*. New York, USA: John Wiley & Sons, Inc.

Yan, X. and X. G. Su (2009). *Linear Regression Analysis: Theory and Computing*. Singapore: World Scientific Publishing Co. Pvt. Ltd.